Journal of Clinical Epidemiology 181 (2025) 111732

ORIGINAL RESEARCH

Assessing the quality of prediction models in health care using the Prediction model Risk Of Bias ASsessment Tool (PROBAST): an evaluation of its use and practical application

Tabea Kaul[a], Johanna A.A. Damen[a,b,*], Laure Wynants[c,d,e], Ben Van Calster[d,e], Maarten van Smeden[a], Lotty Hooft[a,b], Karel G.M. Moons[a]

[a]*Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht University, Utrecht, The Netherlands*
[b]*Cochrane Netherlands, University Medical Center Utrecht, Utrecht University, Utrecht, The Netherlands*
[c]*Department of Epidemiology, CAPHRI Care and Public Health Research Institute, Maastricht University, Maastricht, The Netherlands*
[d]*Department of Development and Regeneration, KU Leuven, Leuven, Belgium*
[e]*Leuven Unit for Health Technology Assessment Research (LUHTAR), KU Leuven, Leuven, Belgium*

Accepted 18 February 2025; Published online 25 February 2025

Abstract

Background and Objectives: Since 2019, the Prediction model Risk Of Bias ASsessment Tool (PROBAST; www.probast.org) has supported methodological quality assessments of prediction model studies. Most prediction model studies are rated with a "High" risk of bias (ROB) and researchers report low interrater reliability (IRR) using PROBAST. We aimed to (1) assess the IRR of PROBAST ratings between assessors of the same study and understand reasons for discrepancies, (2) determine which items contribute most to domain-level ROB ratings, and (3) explore the impact of consensus meetings.

Study Design and Setting: We used PROBAST assessments from a systematic review of diagnostic and prognostic COVID-19 prediction models as a case study. Assessors included international experts in prediction model studies or their reviews. We assessed IRR using prevalence-adjusted bias-adjusted kappa (PABAK) before consensus meetings, examined bias ratings per domain-level ROB judgments, and evaluated the impact of consensus meetings by identifying rating changes after discussion.

Results: We analyzed 2167 PROBAST assessments from 27 assessor pairs covering 760 prediction models: 384 developments, 242 validations, and 134 mixed assessments (including both). The IRR using PABAK was higher for overall ROB judgments (development: 0.82 [0.76; 0.89]; validation: 0.78 [0.68; 0.88]) compared to domain- and item-level judgments. Some PROBAST items frequently contributed to domain-level ROB judgments, eg, 3.5 Outcome blinding and 4.1 Sample size. Consensus discussions mainly led to item-level and never to overall ROB rating changes.

Conclusion: Within this case study, PROBAST assessments received high IRR at the overall ROB level, with some variation at item- and domain-level. To reduce variability, PROBAST assessors should standardize item- and domain-level judgments and hold well-structured consensus meetings between assessors of the same study. © 2025 The Author(s). Published by Elsevier Inc. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

Keywords: Prediction models; Risk of bias assessments; PROBAST; Interrater reliability; Systematic reviews; Methodology

Funding: This study did not receive any funding.

* Corresponding author. Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, P.O. Box 85500, Utrecht 3508 GA, The Netherlands.

E-mail address: j.a.a.damen@umcutrecht.nl (J.A.A. Damen).

https://doi.org/10.1016/j.jclinepi.2025.111732
0895-4356/© 2025 The Author(s). Published by Elsevier Inc. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

Plain Language Summary

The Prediction model Risk Of Bias ASsessment Tool (PROBAST; www.probast.org) provides a set of items to assess the quality of medical studies on so-called prediction tools that calculate an individual's probability of having or developing a certain disease or health outcome. Previous research found low interrater reliability (IRR; ie, how consistently two assessors rate aspects of the same study) when using PROBAST. To understand why this is the case, we conducted a large study involving more than 30 experts from around the world, all of whom applied PROBAST to the same set of prediction tool studies. Based on more than 2150 PROBAST assessments, we identified which PROBAST items led to the most disagreements between raters, explored reasons for these disagreements, and examined whether the use of so-called consensus meetings (ie, different assessors of the same study discuss their ratings and decide on a finalized rating) impacted PROBAST ratings. Our study found that the IRR between different assessors of the same study was higher than previously reported. One explanation for the better agreement compared to previous research may be the preplanning on how to assess certain PROBAST aspects before starting the assessments, as well as holding well-structured consensus meetings. These improvements lead to a more effective use of PROBAST in evaluating the trustworthiness and quality of prediction tools in the health-care domain.

1. Introduction

The Prediction model Risk Of Bias ASsessment Tool (PROBAST; www.probast.org) was launched in 2019 and was designed to support researchers in judging the risk of bias (ROB) and applicability of diagnostic and prognostic prediction models [1,2]. Bias refers to a systematic error in the estimators of the model's predictive performance for the target population of interest and may act in either direction such that there may be over − or under estimation of the true model performance. PROBAST consists of four domains with a total of 20 signaling questions concerning ROB and applicability [1,2]. These signaling questions are rated with ''No'' and ''Probably No'' to flag the potential for bias introduced by specific methodological aspects, ''Yes'' and ''Probably Yes'' indicating no evidence for bias, and ''No Information'' flagging insufficient information for adequate judgment [1,2]. Responses to signaling questions guide domain-level ROB judgments (''High'', ''Low'', and ''Unclear''), which then inform the overall ROB rating of a study.

Most prediction model studies assessed in systematic reviews and other appraisals (eg, for developing medical guidelines) receive ''High'' ROB ratings across different PROBAST domains [3−5]. For example, this may be due to bias introduced by participant selection (ie, PROBAST domain 1) or the analytical methods used to develop or validate the prediction model (ie, PROBAST domain 4) [1,2]. A recent systematic review on prediction models regarding COVID-19 diagnosis and prognosis illustrates this issue: 545 of the 731 assessed prediction models received a ''High'' overall ROB judgment using PROBAST [3].

Besides methodological concerns in prediction models and prediction model studies, other reasons for ''High'' ROB ratings using PROBAST have been explored. Langenhuijsen et al described ''ceiling effects'', ie, the inability to distinguish between degrees of ROB ratings [5]. That is, if a single signaling question receives a ''No'' rating, the domain may be rated as ''High'' ROB; similarly, if all signaling questions receive a ''No'' rating, the same applies [5]. PROBAST guidance leaves whether a single ''No'' should lead to a ''High'' domain-level ROB rating up to the assessors and their specific context [1,2]. Therefore, determining if a ''High'' ROB judgment resulted from just a single signaling question is challenging without knowing how PROBAST assessments were operationalized [1,2,5]. Conversely, PROBAST guidance for overall study ROB judgment are clear: if one ROB domain is rated ''High'', the overall study ROB should be ''High''; ''Low'' overall study ROB should only follow if all domains are rated ''Low'' [1,2].

Several studies have reported poor interrater agreement and reliability among assessors using PROBAST [5,6]. Subjectivity and rater dependency are known pitfalls of any ROB assessment (eg, for randomized intervention trials, diagnostic test accuracy studies, and etiologic studies) and are not limited to PROBAST and prediction model studies [7,8]. Providing training sessions prior to using PROBAST showed improved agreement on domain- and overall-prediction model study ROB judgments [6]. As applies to all types and settings of ROB assessments, to prevent low interrater agreements and reliability, adequate guidance and trainings may be advisable before applying any ROB tool.

We conducted a comprehensive study using a case study of 2167 PROBAST assessments to 1) assess IRR and agreement in PROBAST assessments by identifying signaling questions with significant variation in ratings between different assessors of the same study and understanding the reasons for these discrepancies, 2) determine which PROBAST signaling questions contribute most significantly to domain-level ROB ratings, and 3) explore

What is new?

Key findings

- PROBAST assessments of different assessors of the same study vary on item- and domain-level but agreement is high on overall prediction model risk of bias judgments.
- Some PROBAST items frequently contribute to domain-level risk of bias judgments, while others are sometimes overruled.
- Consensus meetings between different assessors of the same study especially impact PROBAST item- and domain-level ratings, but not overall prediction model risk of bias judgments.

What this adds to what was known?

- Risk of bias assessments suffer from subjectivity and rater dependency.

What is the implication and what should change now?

- To reduce variability between assessors during methodological quality assessments, assessors should standardize item-level ratings and domain-level risk of bias judgments, and hold well-structured consensus meetings between different assessors of the same study.

whether the use of consensus meetings between different assessors of the same study impacts PROBAST ratings.

2. Methods

2.1. Case study

The case study comprised 2167 PROBAST assessments from a systematic review on prediction models for diagnosis and prognosis of COVID-19 [3]. The review included the judgments of 27 pairs of assessors, who were international experts in prediction model studies or their reviews and had varying levels of experience using PROBAST [3]. Each prediction model study was judged by two assessors using PROBAST. Details of the case study, as well as the instructions provided to assessors are part of the Supplementary (section A and B).

2.2. Dataset description

The case study's data extraction sheet comprised two subgroups of data: the individual assessor ratings (where available) as well as the consensus ratings (ie, the agreed ratings between the assessors after discussion). Each row in the data extraction sheet represented the assessment either of a single prediction model study (development or validation procedure), or of a single model from within a study including multiple models/procedures, or of both the model development and validation in one row if assessed together (ie, hereafter referred to as ''mixed assessments'').

We specifically focused on the ROB assessments using PROBAST, including responses to all 20 signaling questions, set into the following four domains: *Participants*, *Predictors*, *Outcome*, and *Analysis*. PROBAST consists of the following two parts: 20 signaling questions for model development and 17 for model validation assessments. At the end, PROBAST includes an overall judgment on the ROB of the prediction model study (hereafter ''overall ROB judgment'') [1,2]. *Model development* refers to the actual development (or training) of the prediction model. *Model validation* refers to evaluating the predictive performance of a model using other data (external validation) [1,2].

Abbreviations for the PROBAST signaling questions (referred to as items) can be found in Table 1.

2.3. Statistical analysis

Prior to all analyses and in accordance with the original guidance of PROBAST [1,2], ratings of ''Probably Yes'' were merged with ''Yes'', and ''Probably No'' with ''No'', as they are treated similarly in PROBAST. All statistical analyses were performed using R Studio (version 2023.12.0 + 369) [9].

2.3.1. Identifying variations in item- and domain-specific interrater reliability and agreement

We calculated the prevalence-adjusted bias-adjusted kappa (PABAK) to evaluate the IRR between different assessors of the same studies [10,11]. Though widely used, Cohen's kappa may underestimate agreement in case of prevalence imbalances [10]. Given the high proportion of ''High'' ROB ratings identified in the case study, we focused our IRR assessments on PABAK but calculated the Cohen's kappa for reference [3,11].

We conducted these analyses for item-level ratings, as well as domain-level and overall ROB judgments. PABAK and Cohen's kappa values were classified as less than 0.21 indicating ''slight'', 0.21−0.40 indicating ''fair'', 0.41−0.60 indicating ''moderate'', 0.61−0.80 indicating ''substantial'', and more than 0.81 indicating ''almost perfect'' agreement. For the absolute agreement between the assessors, we calculated counts and proportions. Incomplete sets of assessments (ie, where the rating of both or one of the assessors was missing) were excluded from the analyses.

Table 1. Abbreviations of PROBAST items

Domains	Abbreviations
1. Participants	
1. Were appropriate data sources used, eg, cohort, RCT, or nested case-control study data?	1.1 Data source
2. Were all inclusions and exclusions of participants appropriate?	1.2 In/exclusion criteria
2. Predictors	
1. Were predictors defined and assessed in a similar way for all participants?	2.1 Predictor standardization
2. Were predictor assessments made without knowledge of outcome data?	2.2 Predictor blinding
3. Are all predictors available at the time the model is intended to be used?	2.3 Predictor availability
3. Outcome	
1. Was the outcome determined appropriately?	3.1 Outcome misclassification
2. Was a prespecified or standard outcome definition used?	3.2 Outcome prespecification
3. Were predictors excluded from the outcome definition?	3.3 Incorporation
4. Was the outcome defined and determined in a similar way for all participants?	3.4 Outcome standardization
5. Was the outcome determined without knowledge of predictor information?	3.5 Outcome blinding
6. Was the time interval between predictor assessment and outcome determination appropriate?	3.6 Time interval
4. Analysis	
1. Were there a reasonable number of participants with the outcome?	4.1 Sample size
2. Were continuous and categorical predictors handled appropriately?	4.2 Handling predictors
3. Were all enrolled participants included in the analysis?	4.3 Missing data
4. Were participants with missing data handled appropriately?	4.4 Missing data handling
5. Was selection of predictors based on univariable analysis avoided?[a]	4.5 Univariable predictor selection
6. Were complexities in the data (eg, censoring, competing risks, and sampling of control participants) accounted for appropriately?	4.6 Data complexity
7. Were relevant model performance measures evaluated appropriately?	4.7 Performance measures
8. Were model overfitting and optimism in model performance accounted for?[a]	4.8 Overfitting
9. Do predictors and their assigned weights in the final model correspond to the results from the reported multivariable analysis?[a]	4.9 Predictor weights

RCT, randomized controlled trial.
[a] Model development assessments only.

2.3.2. Item-level ratings and domain-level ROB judgments

To evaluate how item-level ratings contribute to domain-level ROB judgments, we analyzed "No" responses to the PROBAST items. We specifically examined "No" responses since these indicate the potential for bias and thus contribute to "High" domain-level ROB ratings (Supplementary, section B) [1,2].

For our analysis, we considered a single PROBAST item to be "overruled" to inform domain-level ROB judgments if an item included a "No" response despite a "Low" or "Unclear" domain-level ROB judgment. Accordingly, an item contributed less to domain-level ROB judgments if it was frequently overruled (ie, a high proportion of "No" responses despite a "Low" or "Unclear" domain-level ROB rating). Consequently, an item contributed more to domain-level ROB judgments if it had a low proportion of "No" responses in "Low" or "Unclear" domain-level ROB judgments but high proportions of "No" responses in "High" ROB domain-level judgments.

Besides, we assessed how many studies received at least one "No" response per domain-level ROB rating, to explore whether receiving a "No" response contributed to domain-level ROB judgments (giving an indication whether rating instructions were followed: *Did all domains rated with "High" ROB include at least one "No" response?).*

We calculated the counts and proportions of PROBAST items rated "No" per domain-level ROB rating. Consensus ratings were used for this analysis as they represented the most complete set and ensured a consistent interpretation of the scoring instructions.

2.3.3. Impact of consensus meetings on PROBAST ratings

To explore the impact of consensus meetings between different assessors of the same study on their PROBAST

assessments, we evaluated 1) patterns of rating changes in PROBAST assessments from the assessor to the consensus rating and 2) whether an item-level rating change led to a domain-level ROB judgment change (ie, domain-level shift). A *rating change* was present if both ratings of the assessors of the same study differed from the consensus rating. For example, if assessor 1 scored item 1.1 Data source as ''No Information'' and assessor 2 rated it as ''No'', but after discussion during the consensus meeting, they agreed to rate it as ''Yes''; this represents a rating change. Similarly, if both assessor 1 and 2 initially rated the item as ''No'' but the consensus rating was ''Yes'', a change in rating occurred. Furthermore, a domain-level shift occurred if, eg, assessor 1 rated item 1.1 data source as ''No Information'' and assessor 2 rated 1.1. as ''No'', but the consensus rating included a ''Yes'', this may result in a change on the domain-level. In this case, the domain-level ROB judgment may change from ''Unclear'' (assessor 1) and ''High'' (assessor 2) to a consensus rating of ''Low'' ROB.

We only considered complete assessment sets for the analysis, comprising two individual assessor ratings and one consensus rating per model. To identify patterns of rating changes, we calculated counts and proportions, with the numerator being the number of rating changes and the denominator being the number of models with complete rating sets. To evaluate domain-level shifts, we counted the number of item-level changes per domain and per domain-level changes.

3. Results

Our analysis included 2167 assessments comprising 1407 individual assessor and 760 consensus ratings (for 113 consensus ratings, the individual assessments of one or both assessors were missing). The assessments comprised 760 different diagnostic or prognostic COVID-19 prediction models, including 384 model developments, 242 model validations, as well as 134 mixed assessments. Mixed assessments included model development and validation assessments in one data extraction row, which could not be separated out. These mixed assessments were analyzed separately, and the Supplementary (section C) presents their results.

3.1. Identifying variations in item- and domain-specific interrater reliability and agreement

Table 2 shows the IRR and agreement between the assessors of the same study using PABAK. Most model development items showed ''moderate'' agreement, while most model validation items included ''substantial'' agreement. The overall ROB judgments for model development included ''almost perfect'' agreement and ''substantial'' agreement for model validations. In both model development and validation assessments, agreements were generally lower on domain-level compared to overall ROB judgments. For model development, agreements ranged from ''fair'' (*Outcome*) to ''substantial'' (*Analysis*). In model validation assessments, agreements ranged from ''moderate'' (*Predictors*) to ''substantial'' (*Participants & Analysis*).

On item-level, three of the items included ''slight'' agreement, ie, item 4.6 Data complexity in both model development and validation, as well as item 3.5 Outcome blinding in model validation. Six items had ''fair'' agreement, five in model development and one in model validation (ie, 4.4 Missing data handling). ''Moderate'' agreement was identified for 12 items in model development and four in model validation (ie, 2.1 Predictor standardization, 2.3 Predictor availability, 3.6 Time interval, and 4.2 Handling predictors). Furthermore, ''substantial'' agreement was identified in a total of nine items (two in model development [ie, 4.1 Sample size & 4.7 Univariable predictor selection]; seven in model validation). Only model validation items included ''Almost perfect'' agreement: 1.1 Data source, 3.3. Incorporation, and 4.1 Sample size.

The Supplementary (section D) includes the IRR based on Cohen's kappa. Cohen's kappa and PABAK showed similar IRR trends in model development assessments (both ''moderate'') but lower for model validation assessments (PABAK: ''substantial''; Cohen's kappa: ''fair''). Furthermore, model developments included lower overall ROB agreements using Cohen's kappa compared to PABAK (Cohen's kappa: ''fair''; PABAK: ''almost perfect''), but similar agreements in model validation assessments (both showing ''substantial'' agreement).

3.1.1. Item-level ratings and domain-level ROB judgments

Table 3 shows the proportion of items scored with ''No'' per domain-level ROB judgment. Section E of the supplement includes the number of assessments per domain-level ROB rating with at least one item scored ''No''.

The Participants domain had a ''High'' ROB in 29% (103/355) of model development and 15% (36/235) of model validation assessments. The highest ''No'' ratings in ''High'' ROB judgments were found for item 1.2 In/exclusion criteria (62% for model development; 83% for model validation). While none of the model development assessments included a ''No'' rating in the ''Low'' ROB ratings for item 1.2, one item received a ''No'' response among all model validation assessments (1/164). In addition, 94% of ''High'' ROB model development assessments and all model validation assessments had at least one ''No'' response (Supplementary, section E).

The Predictors domain had a ''High'' ROB in 11% (41/360) of model development and 1% (2/233) of model validation assessments. For model developments, item 2.1 Predictor standardization received a ''No''

Table 2. Interrater reliability (PABAK) and agreement between different assessors of the same prediction model study

	Development			
Domains	**Models n[a]**	**Agreement %**	**PABAK**	**Interpretation**
Domain 1: Participants				
1.1 Were appropriate data sources used, eg, cohort, RCT, or nested case-control study data?	284	74	0.47 (0.37; 0.57)	Moderate
1.2 Were all inclusions and exclusions of participants appropriate?	288	65	0.30 (0.19; 0.41)	Fair
Risk of bias introduced by participants or data sources	288	66	0.31 (0.2; 0.42)	Fair
Domain 2: Predictors				
2.1 Were predictors defined and assessed in a similar way for all participants?	285	75	0.51 (0.41; 0.61)	Moderate
2.2 Were predictor assessments made without knowledge of outcome data?	285	68	0.36 (0.25; 0.47)	Fair
2.3 Are all predictors available at the time the model is intended to be used?	285	76	0.53 (0.43; 0.63)	Moderate
Risk of bias introduced by predictors or their assessment	287	64	0.28 (0.16; 0.39)	Fair
Domain 3: Outcome				
3.1 Was the outcome determined appropriately?	284	77	0.55 (0.45; 0.65)	Moderate
3.2 Was a prespecified or standard outcome definition used?	287	79	0.57 (0.48; 0.67)	Moderate
3.3 Were predictors excluded from the outcome definition?	286	80	0.59 (0.50; 0.69)	Moderate
3.4 Was the outcome defined and determined in a similar way for all participants?	288	74	0.49 (0.39; 0.59)	Moderate
3.5 Was the outcome determined without knowledge of predictor information?	287	62	0.24 (0.13; 0.35)	Fair
3.6 Was the time interval between predictor assessment and outcome determination appropriate?	287	72	0.44 (0.34; 0.55)	Moderate
Risk of bias introduced by the outcome or its determination	287	61	0.23 (0.11; 0.34)	Fair
Domain 4: Analysis				
4.1 Were there a reasonable number of participants with the outcome?	285	82	0.64 (0.55; 0.72)	Substantial
4.2 Were continuous and categorical predictors handled appropriately?	286	73	0.45 (0.35; 0.56)	Moderate
4.3 Were all enrolled participants included in the analysis?	284	68	0.37 (0.26; 0.47)	Fair
4.4 Were participants with missing data handled appropriately?	284	76	0.53 (0.43; 0.63)	Moderate
4.5 Was selection of predictors based on univariable analysis avoided?	286	85	0.70 (0.62; 0.78)	Substantial
4.6 Were complexities in the data (eg, censoring, competing risks, sampling of control participants) accounted for appropriately?	281	58	0.16 (0.04; 0.28)	Slight
4.7 Were relevant model performance measures evaluated appropriately?	286	74	0.48 (0.38; 0.58)	Moderate
4.8 Were model overfitting and optimism in model performance accounted for?	282	68	0.35 (0.25; 0.46)	Fair
4.9 Do predictors and their assigned weights in the final model correspond to the results from the reported multivariable analysis?	283	74	0.48 (0.38; 0.59)	Moderate
Risk of bias introduced by the analysis	287	84	0.68 (0.59; 0.76)	Substantial
Overall				
Reaching an overall judgment about risk of bias of the prediction model	281	91	0.82 (0.76; 0.89)	Almost perfect

RCT, randomized controlled trial.
PABAK, prevalence-adjusted bias-adjusted kappa (<0.21: slight, 0.21–0.40: fair, 0.41–0.60: moderate, 0.61–0.80: substantial, 0.81–0.99: almost perfect, and 1: perfect).
Prior to all analyses, the item-level rating ''Probably No'' was merged with ''No'', while ''Probably Yes'' was merged with ''Yes''.
[a] Number of models with individual and consensus ratings; IRR, Interrater reliability.
[b] Missing: incomplete sets of assessments (ie, where the rating of both or one of the assessors was missing).

rating in 54% of assessments including a ''High'' ROB rating, while none did in ''Low'' ROB ratings. For model validation assessments, item 2.1 and item 2.2 Predictor blinding received ''No'' responses in 50% of the assessments (1 out of 2). Item 2.2 Predictor blinding never received a ''No'' response in assessments with ''Low'' ROB ratings. For ''High'' ROB ratings, 90% (37/41) of model developments included at least one

Missing[b] %	Validation				Missing[b] %
	Models n[a]	Agreement %	PABAK	Interpretation	
26	149	91	0.81 (0.72; 0.91)	Almost perfect	38
25	149	89	0.77 (0.67; 0.87)	Substantial	38
25	149	88	0.76 (0.65; 0.86)	Substantial	38
26	147	78	0.55 (0.42; 0.69)	Moderate	39
26	147	88	0.76 (0.65; 0.86)	Substantial	39
26	147	78	0.56 (0.43; 0.70)	Moderate	39
25	147	73	0.47 (0.33; 0.61)	Moderate	39
26	148	84	0.69 (0.57; 0.81)	Substantial	39
25	148	89	0.77 (0.67; 0.87)	Substantial	39
26	148	92	0.84 (0.75; 0.93)	Almost perfect	39
25	148	86	0.73 (0.62; 0.84)	Substantial	39
25	147	58	0.16 (0; 0.32)	Slight	39
25	148	78	0.55 (0.42; 0.69)	Moderate	39
25	148	84	0.69 (0.57; 0.81)	Substantial	39
26	149	97	0.93 (0.88; 0.99)	Almost perfect	38
26	141	75	0.50 (0.36; 0.65)	Moderate	42
26	147	83	0.66 (0.54; 0.78)	Substantial	39
26	150	64	0.28 (0.13; 0.43)	Fair	38
26	-	-	-	-	-
27	141	52	0.04 (−0.13; 0.20)	Slight	42
26	150	84	0.68 (0.56; 0.80)	Substantial	38
27	-	-	-	-	-
26	-	-	-	-	-
25	141	88	0.76 (0.65; 0.87)	Substantial	42
27	146	89	0.78 (0.68; 0.88)	Substantial	40

''No'' response, as did 50% of model validations (1/2) (Supplementary, section E).

The Outcome domain included ''High'' ROB in 19% (70/360) of the model development and 14% (34/237) of model validation assessments. The highest ''No'' ratings were found for item 3.5 Outcome blinding for the ''High'' ROB judgments (43% for model development; 76% for model validation). However, both included some ''No''

ratings in the ''Low'' ROB ratings. Interestingly, 86% of model development assessments judged with a ''High'' ROB included at least one item rated with ''No'', while 97% of model validation assessments did (Supplementary, section E).

In the Analysis domain, 90% (314/351) of model development and 79% (178/226) of model validation assessments included ''High'' ROB judgments. Item 4.1 Sample size included the highest proportion of ''No'' ratings for assessments including a ''High'' ROB judgment (both 73%), while also including one ''No'' rating in the ''Low'' ROB judgments (14%; 4%). For ''High'' ROB ratings, 95% of model developments, and 81% of model validations included at least one item with a ''No'' response (Supplementary, section E).

3.2. Impact of consensus meetings on PROBAST ratings

3.2.1. Patterns of rating changes

Table 4 shows the rating changes from the two individual assessors to the consensus rating. Rating changes ranged from 0% to 6%. The model development assessments included more rating changes compared to the model validation assessments. The Analysis domain included most rating changes, particularly in item 4.6 Data complexity (6%). However, fewer changes were identified on domain-level and none in the overall ROB judgments.

3.2.2. Domain-level shifts

The number of domain-level shifts ranged from none (Analysis domain in model validation assessments) to 10 (Participants domain in model development assessments) (Supplementary, section F). Most item-level rating changes occurred in model development assessments; however, relative to the number of item-level rating changes, most domain-level shifts were observed in model validation assessments (Supplementary, section F).

4. Discussion

We performed a comprehensive study on the use and practical application of the Prediction model Risk Of Bias ASsessment Tool (PROBAST; www.probast.org) [1,2] to assess diagnostic and prognostic prediction models using a case study of a large systematic review on COVID-19 prediction models [3]. Our results show that the IRR using PROBAST, based on the prevalence-adjusted bias-adjusted kappa (PABAK), was high for overall ROB judgments but lacks agreement on some item-level ratings and domain-level ROB judgments. Furthermore, IRR was generally higher in prediction model validation assessments compared to prediction model development assessments. We identified certain PROBAST items that frequently contributed to domain-level ROB judgments. The use of consensus meetings between different assessors of the same study impacted item-level ratings to a larger extent than domain-level ROB judgments, while not showing any impact on overall ROB judgments.

4.1. Comparison to other studies

Our results partly reflect findings reported by others [5,6,12]. We observed varying IRR across PROBAST assessments but cannot confirm findings of consistently low IRR [5,6]. Use of the PABAK as opposed to the Cohen's kappa and differences in PROBAST operationalization (ie, how domains should be rated) may be possible explanations for these variations [4,10,13].

We identified varying trends in IRR between model development and validation assessments. This distinction is often overlooked in prior research, where IRR calculations frequently combine data from various systematic reviews without distinguishing between these evaluation types [5,6]. Furthermore, due to the difference in assessors and research teams, the assessments may have been conducted with a different set of instructions (ie, operationalizations), leading to greater variability in responses (ie, lower IRR) [13]. In our study, the assessors received the same set of instructions and consensus meetings were planned (per protocol) throughout the five assessment rounds to finalize ratings. Thus, subjectivity could be limited.

We found higher IRR between different assessors of the same study compared to previous reports [5,6]. However, while we based our main IRR assessments on the PABAK, previous studies frequently used the Cohen's kappa. Cohen's kappa may underestimate IRR when a certain category is overrepresented (eg, as in our case, a large proportion of ''High'' ROB ratings for most assessments of the Analysis domain) [3,10]. For reference, we included the IRR based on Cohen's kappa in the Supplementary (section D). While the Cohen's kappa showed similar trends as by PABAK for some PROBAST items, others included lower IRR. Therefore, focusing on the PABAK to assess IRR seemed appropriate.

Adequate instructions and operationalization prior to any ROB assessments are key to ensuring consistency and reliability across assessments [7,14,15]. In our study, we identified certain PROBAST items that frequently contributed to domain-level ROB judgments, while others were frequently ''overruled''. However, the instructions given in our case study may differ from those of other PROBAST assessments. It is difficult to evaluate the differences in how domain-level ROB judgments were operationalized across different reports without understanding the specific criteria assessors used. To ensure comparability of methodological quality assessments, reporting checklists (eg, Preferred Reporting Items for Systematic reviews and Meta-Analysis [PRISMA]) might consider including operationalizations as a reporting item [16].

Table 3. Number of items scored ''No'' per domain-level ROB judgment[a]

	Development *n* (%)[b]			Validation *n* (%)[b]		
Domains	**High**	**Low**	**Unclear**	**High**	**Low**	**Unclear**
Participants	***N* = 103**	***N* = 146**	***N* = 106**	***N* = 36**	***N* = 164**	***N* = 35**
1. Were appropriate data sources used, eg, cohort, RCT, or nested case-control study data?	56 (54)	4 (3)	2 (2)	10 (28)	0 (0)	0 (0)
2. Were all inclusions and exclusions of participants appropriate?	64 (62)	0 (0)	5 (5)	30 (83)	1 (1)	10 (29)
Predictors	***N* = 41**	***N* = 188**	***N* = 131**	***N* = 2**	***N* = 160**	***N* = 71**
1. Were predictors defined and assessed in a similar way for all participants?	22 (54)	0 (0)	2 (2)	1 (50)	7 (4)	0 (0)
2. Were predictor assessments made without knowledge of outcome data?	13 (32)	4 (2)	3 (2)	1 (50)	0 (0)	0 (0)
3. Are all predictors available at the time the model is intended to be used?	5 (12)	4 (2)	1 (1)	0 (0)	0 (0)	0 (0)
Outcome	***N* = 70**	***N* = 140**	***N* = 150**	***N* = 34**	***N* = 173**	***N* = 30**
1. Was the outcome determined appropriately?	16 (23)	0 (0)	3 (2)	13 (38)	0 (0)	0 (0)
2. Was a prespecified or standard outcome definition used?	12 (17)	0 (0)	3 (2)	10 (29)	0 (0)	0 (0)
3. Were predictors excluded from the outcome definition?	18 (26)	4 (3)	1 (1)	16 (47)	1 (1)	3 (10)
4. Was the outcome defined and determined in a similar way for all participants?	22 (31)	0 (0)	7 (5)	12 (35)	0 (0)	0 (0)
5. Was the outcome determined without knowledge of predictor information?	30 (43)	9 (6)	14 (9)	26 (76)	7 (4)	4 (13)
6. Was the time interval between predictor assessment and outcome determination appropriate?	14 (20)	3 (2)	2 (1)	11 (32)	0 (0)	4 (13)
Analysis	***N* = 314**	***N* = 7**	***N* = 30**	***N* = 178**	***N* = 26**	***N* = 22**
1. Were there a reasonable number of participants with the outcome?	229 (73)	1 (14)	11 (37)	130 (73)	1 (4)	5 (23)
2. Were continuous and categorical predictors handled appropriately?	114 (36)	0 (0)	2 (7)	19 (11)	0 (0)	0 (0)
3. Were all enrolled participants included in the analysis?	80 (25)	3 (43)	1 (3)	80 (45)	3 (12)	10 (45)
4. Were participants with missing data handled appropriately?	121 (39)	2 (29)	2 (7)	87 (49)	3 (12)	8 (36)
5. Was selection of predictors based on univariable analysis avoided?	105 (33)	0 (0)	3 (10)	-	-	-
6. Were complexities in the data (eg, censoring, competing risks, and sampling of control participants) accounted for appropriately?	103 (33)	0 (0)	0 (0)	35 (20)	2 (8)	6 (27)
7. Were relevant model performance measures evaluated appropriately?	191 (61)	1 (14)	9 (30)	92 (52)	1 (4)	10 (45)
8. Were model overfitting and optimism in model performance accounted for?	204 (65)	1 (14)	7 (23)	-	-	-
9. Do predictors and their assigned weights in the final model correspond to the results from the reported multivariable analysis?	22 (7)	0 (0)	1 (3)	-	-	-

RCT, randomized controlled trial.

Prior to all analyses, the item-level rating ''Probably No'' was merged with ''No'', while ''Probably Yes'' was merged with ''Yes''.

[a] The first row for each domain includes the number of assessments with a specific domain ROB rating (horizontally), following the number of items per domain ROB rating judged with ''No'' (vertically).

[b] Proportion: numerator: number of items scored ''No'' per domain-level ROB judgment; denominator: number of models per domain-level ROB judgment.

Table 4. Rating changes from individual to consensus ratings

Domains	Development		Validation	
	n (%)[a]	Models n[b]	*n* (%)[a]	Models n[b]
Participants				
1. Were appropriate data sources used, eg, cohort, RCT, or nested case-control study data?	2 (1)	252	0 (0)	149
2. Were all inclusions and exclusions of participants appropriate?	9 (2)	255	3 (1)	149
Risk of bias introduced by participants or data sources	15 (4)	265	4 (2)	149
Predictors				
1. Were predictors defined and assessed in a similar way for all participants?	6 (2)	253	2 (1)	147
2. Were predictor assessments made without knowledge of outcome data?	2 (1)	253	0 (0)	147
3. Are all predictors available at the time the model is intended to be used?	1 (0)	252	2 (1)	147
Risk of bias introduced by predictors or their assessment	13 (3)	270	4 (2)	147
Outcome				
1. Was the outcome determined appropriately?	2 (1)	252	0 (0)	148
2. Was a prespecified or standard outcome definition used?	3 (1)	254	1 (0)	148
3. Were predictors excluded from the outcome definition?	3 (1)	252	0 (0)	148
4. Was the outcome defined and determined in a similar way for all participants?	6 (2)	256	1 (0)	148
5. Was the outcome determined without knowledge of predictor information?	3 (1)	256	1 (0)	147
6. Was the time interval between predictor assessment and outcome determination appropriate?	4 (1)	255	0 (0)	148
Risk of bias introduced by the outcome or its determination	6 (2)	268	1 (0)	148
Analysis				
1. Were there a reasonable number of participants with the outcome?	20 (5)	250	1 (0)	149
2. Were continuous and categorical predictors handled appropriately?	21 (5)	252	2 (1)	140
3. Were all enrolled participants included in the analysis?	18 (5)	250	1 (0)	146
4. Were participants with missing data handled appropriately?	13 (3)	250	1 (0)	149
5. Was selection of predictors based on univariable analysis avoided?	16 (4)	250	-	-
6. Were complexities in the data (eg, censoring, competing risks, sampling of control participants) accounted for appropriately?	22 (6)	247	6 (2)	141
7. Were relevant model performance measures evaluated appropriately?	17 (4)	251	1 (0)	150
8. Were model overfitting and optimism in model performance accounted for?	16 (4)	245	-	-
9. Do predictors and their assigned weights in the final model correspond to the results from the reported multivariable analysis?	15 (4)	250	-	-
Risk of bias introduced by the analysis	9 (2)	261	3 (1)	141
Overall				
Reaching an overall judgment about risk of bias of the prediction model evaluation	0 (0)	254	0 (0)	143

RCT, randomized controlled trial.

Prior to all analyses, the item-level rating ''Probably No'' was merged with ''No'', while ''Probably Yes'' was merged with ''Yes''.

[a] Numerator: number of identified rating changes from individual to consensus rating (ie, both individual model assessments had to differ from the consensus rating); denominator: number of complete assessment sets.

[b] Number of models with a complete assessment set (ie, two individual assessor ratings and one consensus rating).

PROBAST presents a minimal set of methodological quality criteria which merely *guide* the assessments of prediction model studies for diagnosis and prognosis [1,2]. In our study, some PROBAST items frequently contributed to domain-level ROB judgments. Venema et al published a short form of PROBAST in 2021 consisting of six selectively summarized items [12]. Using short forms of PROBAST should be approached with caution, as they may falsely encourage assessors to completely disregard certain aspects of methodological quality (eg, 1.1 Data source or 1.2 In/exclusion criteria [Domain 1: Participants]), which may severely impact a study's validity. This

may especially be the case for inexperienced assessors or when using PROBAST to help guide the design of a prediction model study. PROBAST was developed to thoroughly assess the ROB in the entire prediction model study, addressing all relevant methodological aspects of a prediction model study.

As applies to all ROB tools, consensus meetings between assessors of the same study have a critical role in improving methodological quality assessments. In our study, rating changes occurred on both item- and domain-level judgments after consensus discussion and even in case of previous agreement between the individual assessors. Therefore, holding actual consensus discussions proofed beneficial. These findings highlight the need for structured, well-facilitated consensus meetings between assessors of the same study.

4.2. Implications

To ensure consistency and reliability in ROB assessments, adequate training and standardized assessment procedures are essential. Although assessors were trained experts with varying levels of experience in using PROBAST, our analyses showed that certain PROBAST items would still benefit from more concise explanations and clear operationalization by the users prior to performing PROBAST assessments. Moreover, it is crucial to hold actual consensus meetings during ROB assessments. Even when assessors initially agree on their individual ratings, discussions may still lead to changes in the domain-level ROB judgments.

These improvements have been incorporated into the recent development of PROBAST + AI [17,18] (www.probast.org) which is a continuation and update of PROBAST-2019.

4.3. Limitations

Our study has several limitations. First, our case study included prediction models which were developed and/or validated during the global pandemic COVID-19. Due to the pandemic's urgency, development, validation, and PROBAST assessments were performed in a short amount of time. Furthermore, specific aspects (eg, predicted outcomes) of the case study may have affected our results more than others. For example, in prediction model studies with an objective outcome (case study), such as mortality, PROBAST assessments will differ compared to those done in subjective outcomes, such as quality of life (ie, see item 3.5 Outcome blinding). Therefore, our findings might not reflect PROBAST assessments performed in other medical fields or settings.

Second, for a few items, individual assessments had high percentages of missing values. Though reviewers were asked to submit their individual ratings, this was not always done. It is possible that reviewers preferred to wait for the consensus meeting or consult an expert for difficult items, which may also explain why the consensus rating set was more complete than the individual assessments, and why certain items were missing more frequently than others.

Third, due to the high number of assessments, we were unable to critically evaluate any free-text comments assessors made in detail. Nevertheless, we randomly sampled 10 consensus ratings to explore any information assessors noted down. Notes primarily included reasons for specific domain-level ROB judgments.

Fourth, our results may be influenced by a certain ''learning effect'' that occurred in the multistage review of the case study [3]. However, while some assessors were involved from the beginning, others only joined in later rounds, with varying workloads. Therefore, the ''learning effect'' likely only had a limited impact.

5. Conclusion

For overall ROB judgments, PROBAST assessments include high interrater agreement and reliability. Some variation exists on item- and domain-level assessments. To ensure PROBAST assessments can be performed adequately and coherently, assessors involved in prediction model reviews or guideline development should therefore operationalize on how to judge PROBAST items and domains prior to conducting PROBAST assessments. We also encourage PROBAST users to hold well-structured consensus meetings between assessors of the same study. This will help to ensure that PROBAST assessments are less subjective to variability. The findings of this study have led to the recent development of PROBAST + AI as an extension and update of PROBAST-2019.

CRediT authorship contribution statement

Tabea Kaul: Writing − original draft, Visualization, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Johanna A.A. Damen:** Writing − review & editing, Methodology, Investigation, Formal analysis, Conceptualization. **Laure Wynants:** Writing − review & editing, Methodology, Investigation. **Ben Van Calster:** Writing − review & editing, Methodology, Investigation. **Maarten van Smeden:** Writing − review & editing, Methodology, Investigation. **Lotty Hooft:** Writing − review & editing, Supervision, Methodology, Investigation, Conceptualization. **Karel G.M. Moons:** Writing − review & editing, Supervision, Methodology, Investigation, Conceptualization.

Declaration of competing interest

K.G.M.M. was involved in the development of the quality assessment tool included and described in this study. L.W., B.V.C., K.G.M.M., L.H., M.v.S., and J.A.A.G. were involved in the quality assessment ratings included and

described in our study. There are no competing interests for any other author.

Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jclinepi.2025.111732.

Data availability

Data will be made available on request.

References

[1] Moons KGM, Wolff RF, Riley RD, Whiting PF, Westwood M, Collins GS, et al. PROBAST: a tool to assess risk of bias and applicability of prediction model studies: explanation and elaboration. Ann Intern Med 2019;170(1):W1−33.

[2] Wolff RF, Moons KGM, Riley RD, Whiting PF, Westwood M, Collins GS, et al. PROBAST: a tool to assess the risk of bias and applicability of prediction model studies. Ann Intern Med 2019; 170(1):51−8.

[3] Wynants L, Van Calster B, Collins GS, Riley RD, Heinze G, Schuit E, et al. Prediction models for diagnosis and prognosis of covid-19: systematic review and critical appraisal. BMJ 2020;369: m1328.

[4] Kaiser I, Mathes S, Pfahlberg AB, Uter W, Berking C, Heppt MV, et al. Using the prediction model risk of bias assessment tool (PROBAST) to evaluate melanoma prediction studies. Cancers 2022; 14(12):3033.

[5] Langenhuijsen LFS, Janse RJ, Venema E, Kent DM, van Diepen M, Dekker FW, et al. Systematic meta-review of prediction studies demonstrates stable trends in bias and low PROBAST inter-rater agreement. J Clin Epidemiol 2023;159:159−73.

[6] Kaiser I, Pfahlberg AB, Mathes S, Uter W, Diehl K, Steeb T, et al. Inter-rater agreement in assessing risk of bias in melanoma prediction studies using the prediction model risk of bias assessment tool (PROBAST): results from a controlled experiment on the effect of specific rater training. J Clin Med 2023;12(5):1976.

[7] Tomlinson E, Cooper C, Davenport C, Rutjes AWS, Leeflang M, Mallett S, et al. Common challenges and suggestions for risk of bias tool development: a systematic review of methodological studies. J Clin Epidemiol 2024;171:111370.

[8] Kanukula R, McKenzie JE, Cashin AG, Korevaar E, McDonald S, Mello AT, et al. Variation observed in consensus judgments between pairs of reviewers when assessing the risk of bias due to missing evidence in a sample of published meta-analyses of nutrition research. J Clin Epidemiol 2024;166:111274.

[9] R Core Team. R: A language and environment for statistica computing. R version 2023.12.0+369, Vienna, Austria: R Foundation for Statistical Computing. 2023. Available at: https://www.R-project.org/. Accessed January 31, 2024.

[10] Byrt T, Bishop J, Carlin JB. Bias, prevalence and kappa. J Clin Epidemiol 1993;46(5):423−9.

[11] Cohen J. A coefficient of agreement for nominal scales. Educ Psychol Meas 1960;20(1):37−46.

[12] Venema E, Wessler BS, Paulus JK, Salah R, Raman G, Leung LY, et al. Large-scale validation of the prediction model risk of bias assessment Tool (PROBAST) using a short form: high risk of bias models show poorer discrimination. J Clin Epidemiol 2021;138:32−9.

[13] Hartling L, Hamm MP, Milne A, Vandermeer B, Santaguida PL, Ansari M, et al. Testing the Risk of Bias tool showed low reliability between individual reviewers and across consensus assessments of reviewer pairs. J Clin Epidemiol 2013;66(9):973−81.

[14] Minozzi S, Dwan K, Borrelli F, Filippini G. Reliability of the revised Cochrane risk-of-bias tool for randomised trials (RoB2) improved with the use of implementation instruction. J Clin Epidemiol 2022; 141:99−105.

[15] Crocker TF, Lam N, Jordão M, Brundle C, Prescott M, Forster A, et al. Risk-of-bias assessment using Cochrane's revised tool for randomized trials (RoB 2) was useful but challenging and resource-intensive: observations from a systematic review. J Clin Epidemiol 2023;161:39−45.

[16] Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. Syst Rev 2021;10(1):89.

[17] Collins GS, Dhiman P, Andaur Navarro CL, Ma J, Hooft L, Reitsma JB, et al. Protocol for development of a reporting guideline (TRIPOD-AI) and risk of bias tool (PROBAST-AI) for diagnostic and prognostic prediction model studies based on artificial intelligence. BMJ Open 2021;11(7):e048008.

[18] Andaur Navarro CL, Damen JAA, Hooft L, Reitsma JB, van Smeden M, Riley RD, et al. Protocol for development of a risk of bias assessment tool for diagnostic and prognostic prediction studies based on artificial intelligence: PROBAST-AI 2022. Available at: https://doi.org/10.17605/OSF.IO/4BJWT. Accessed August 25, 2024.